Aplicación de *learning analytics* y *educational data mining* en una institución de educación superior en Colombia[*]

Johnny Salazar-Cardona[**]
Jorge Triviño-Arbeláez[***]

Recibido: 20/10/2018 • Aceptado: 28/06/2019
https://doi.org/10.22395/rium.v19n36a4

Resumen

Actualmente los datos son un insumo clave para el crecimiento continuo de las organizaciones a partir de la toma de decisiones. Crecimiento logrado a través de la aplicación de procesos de descubrimiento de conocimiento, realizando sobre los datos un preprocesamiento, transformación y análisis. El campo académico no es ajeno a este tipo de aplicaciones y es tendencia para el aprovechamiento de los datos generados diariamente sobre sus estudiantes, áreas administrativas y académicas en favor de mejorar continuamente los procesos.

Las metodologías actuales proponen dos directrices: *Learning Analytics* (LA) enfocado principalmente en procesos descriptivos y *Educational Data Mining* (EDM) para procesos predictivos, este direcciona actividades ajustadas a este entorno para obtener resultados satisfactorios. Es por esto que este artículo presenta la aplicación de estas dos corrientes en una institución de educación superior, enfocándolas a datos sensibles de los estudiantes, que servirán de apoyo para la alta directiva de la institución.

Palabras clave: procesos de descubrimiento de conocimiento; *educational data mining; learning analytics;* toma de decisiones; procesos descriptivos; procesos predictivos.

[*] Artículo resultado de una investigación terminada, titulada *Aplicación de un prototipo de analítica visual en el programa de ingeniería de software para apoyar la toma de decisiones*. Dos años de ejecución entre 2016 y 2018. Financiado por la Institución Universitaria EAM.

[**] Magíster en Ingeniería Computacional de la Universidad de Caldas, Manizales. Ingeniero de Sistemas y Computación. Docente de planta de la Institución Universitaria EAM, Armenia, Colombia. Adscrito a la Facultad de Ingeniería y director del grupo de investigación IngeSoft. Correo electrónico: jasalazar@eam.edu.co. Orcid: http://orcid.org/0000-0002-6048-740X

[***] Magíster en Ingeniería de la Universidad Eafit. Ingeniero de sistemas de la Universidad del Quindío, Colombia. Docente e investigador de la Universidad del Quindío, Armenia, Colombia; docente de la Institución Universitaria EAM, Armenia, Colombia. Correo electrónico: jitrivino@uniquindio.edu.co. Orcid: http://orcid.org/0000-0002-1264-3519

Application of learning analytics and educational data mining in an institution of superior education in Colombia

Abstract

Nowadays, data is a key element for the continuous improvement of an organization's decision-taking, achieved through the application of awareness and knowledge processes by undergoing a pre-processing, transformation and analysis over the data. The academic field is aware of this kind of application and is a trend for the exploitation of data generated by the students, its management and academics dependencies on a daily basis in order to continuously improve the processes.

Current methodologies propose two different guidelines: Learning Analytics (LA), primarily focused on descriptive processes, and Educational Data Mining (EDM) for predictive processes, directing activities adjusted to this environment for obtaining satisfactory results. It is for this reason that this article presents an application of these two guidelines in a higher education institution, focusing them on sensitive data of the students that will support the high direction decision-taking in these institutions.

Keywords: Knowledge discovery processes; educational data mining; learning analytics; decision taking; descriptive processes; predictive processes.

Aplicação de *learning analytics* e *educational data mining* em uma instituição de ensino superior na Colômbia

Resumo

Atualmente, os dados são um insumo-chave para o crescimento contínuo das organizações a partir da tomada de decisões, atingido por meio da aplicação de processos de descobrimento de conhecimento, em que se realizam pré-processamento, transformação e análise dos dados. O campo acadêmico não é alheio a esse tipo de aplicações e é tendência para o aproveitamento dos dados gerados diariamente de seus estudantes, áreas administrativas e acadêmicas em prol de melhorar os processos de forma contínua. As metodologias atuais propõem duas diretrizes: *Learning Analytics* focada principalmente em processos descritivos e *Educational Data Mining* para processos preditivos, que direcionam atividades ajustadas a esse contexto para obter resultados satisfatórios. Por isso, este artigo apresenta a aplicação dessas duas correntes em uma instituição de ensino superior, focando-as em dados sensíveis dos estudantes, que servirão de apoio para altas direções na instituição de ensino superior.

Palavras-chave: processos de descobrimento de conhecimento; *educational data mining*; *learning analytics*; tomada de decisões; processos descritivos; processos preditivos.

INTRODUCCIÓN

Las instituciones de educación superior vistas como una organización, intentan continuamente aplicar procesos de mejora, apoyándose en la toma de decisiones por parte de los interesados. Es por esta necesidad que se integró paulatinamente el proceso de análisis de históricos de datos, acoplando el concepto de procesos de descubrimiento de conocimiento en el sector y aplicando un concepto dado en el campo empresarial conocido como inteligencia de negocios (BI) a esta área. En la aplicación de estos elementos en el campo académico se pueden encontrar dos corrientes de investigación que establecen los lineamientos de aplicación de los procesos de descubrimiento de conocimiento: LA *(Learning Analytics)* y EDM *(Educational Data Mining)*, ambos conceptos, aunque similares, poseen elementos clave que los diferencian. En LA el eje principal de aplicación es el ser humano para el análisis de datos y la interpretación de resultados de una manera gráfica y visual, brindando interactividad, gráficas intuitivas, filtros de búsqueda y reportes dinámicos predefinidos para que el analista sea quien emita un juicio final interpretando los resultados. EDM se centra en el análisis de elementos específicos e individuales de manera automatizada, aquí el ser humano no tiene mayor relevancia [1]. Cabe destacar que EDM y LA son corrientes relevantes en el proceso de descubrimiento de conocimiento en datos del sector académico, uno no es mejor o superior al otro, pues su uso radica en el resultado final que se quiere obtener de acuerdo a los enfoques que ofrecen.

En la aplicación de estos enfoques se realizó un proceso de centralización de datos, los cuales fueron tratados computacionalmente. Posteriormente se aplicó un proceso descriptivo con analítica visual, utilizando informes dinámicos predefinidos. Luego se aplicaron procesos predictivos para profundizar aún más en los análisis, en las preguntas que se deseaba responder y los patrones iniciales identificados en los procesos descriptivos. Estos análisis fueron aplicados a los datos de estudiantes de una institución de educación superior, cumpliendo las normativas de protección del *habeas data* [2], y se enfocó en el análisis de diferentes aspectos socioeconómicos y académicos durante los años 2007 al 2017 [3].

1. MATERIALES Y MÉTODOS

Para la aplicación de EDM y LA, es necesario comprender que ambos elementos provienen de un concepto denominado KDP (*Knowledge Discovery Process*), que es una ideología de implementación de tareas y actividades que deben ser ejecutadas para descubrir conocimiento en un conjunto de datos sistematizados. Esta ideología ha tenido diferentes metodologías según el campo donde se desee aplicar: 1) KDD (descubrimiento de conocimiento en bases de datos), es un proceso para descubrir cono-

cimiento útil de un conjunto de datos estructurados o semiestructurados; 2) CRISP-DM (*Cross Industry Standard Process for Data Mining*); 3) Catalyst o P3QT; 4) SEMMA (*Sample – Explore – Model – Modify - Assess*); 5) LA (*Learning Analytics*); 6) EDM (*Educational Data Mining*) entre otros. De estos modelos existentes, el primero que fue definido es KDD en 1996 como un modelo para el campo de la investigación [4-6], y posteriormente lo hicieron con los mencionados anteriormente. Entre los modelos más destacados se encuentran CRISP-DM, especializado en el campo industrial, KDD que ofrece flexibilidad en el proceso de ejecución y EDM-LA para el contexto académico que fueron utilizados para la ejecución de esta investigación.

Respecto a LA y EDM, son comunidades de investigación para la aplicación de KDP a datos de un entorno educativo. De estas comunidades, LA hace uso del concepto de analítica visual para describir el comportamiento de los datos que se generan en un contexto académico. Estos datos involucran elementos como niveles de complejidad de los cursos, identificación de aspectos que afectan el rendimiento académico, recepción positiva o negativa de los estudiantes frente a diferentes temáticas, trazabilidad académica, deserción estudiantil, control de accesos a plataformas virtuales de educación, entre otras. Durante su historia, el término de *Learning Analytics* ha sido debatido frente al concepto y comunidad de *Educational Data Mining* [7-9], con diferencias como el enfoque de la investigación y el tamaño de los datos. Esta brecha entre los enfoques de datos investigados y el tamaño de los *datasets* es actualmente inexistente, debido a que en las diferentes divulgaciones científicas de estas dos comunidades los temas de investigación son similares, aunque es el enfoque aplicado para el análisis de los datos el que marca una diferencia entre estas. LA hace uso del concepto de *Business Intelligence*, específicamente la analítica visual y web, donde el elemento principal de juicio es el ser humano que interpreta y enfoca los análisis según sus necesidades de manera intuitiva realizando análisis descriptivos. Por su parte, EDM se enfoca en la aplicación de elementos predictivos semiautomáticos, donde el ser humano pierde dicha relevancia y no es quien emite de primera mano juicios para la interpretación de resultados [1].

Como se mencionó anteriormente, EDM y LA exponen un proceso de aplicación y metodología de enfoque basadas de la ideología KDP (ver figura 1). En esta LA plantea 3 etapas: 1) *Data collection and pre-processing*, que integra los datos de diferentes fuentes en un único repositorio de datos o *data warehouse,* para posteriormente aplicar sobre estos un conjunto de reglas de negocio que permitirán tener datos adecuados para analizar. Esta etapa es definida a partir de la unión de diferentes tareas de KDP, como el análisis y comprensión del entorno de los datos, creación de la base de datos de trabajo, limpieza y transformación de los datos; 2) *Analytics and action*, es la aplicación de las técnicas de análisis de datos para su interpretación, detección de patrones y a

partir del objetivo del análisis, tomar las medidas respectivas junto con la divulgación de resultados. Esta etapa toma las acciones de comprensión y elección de la técnica de minería de datos, su aplicación, interpretación–evolución y procesamiento de resultados de la ideología KDP; 3) *Post-processing*, enfocado al refinamiento de la base de datos de trabajo con el fin de extender los análisis realizados, añadiendo nuevas fuentes de datos, atributos y extendiendo a nuevos enfoques los análisis previamente realizados.

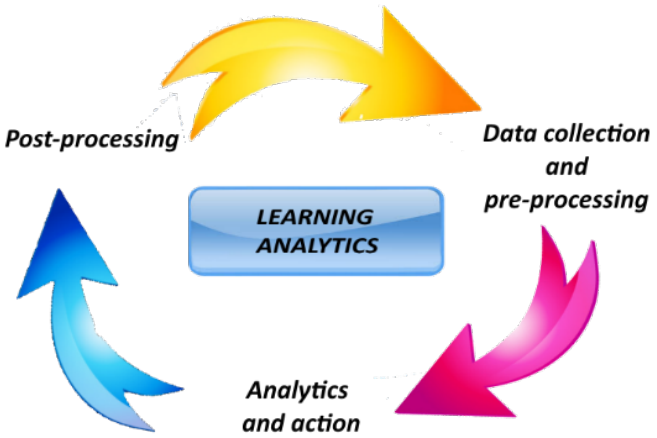

Figura 1. Proceso Learning Analytics
Fuente: A. Chatti, *et al.* [10].

Respecto al modelo de aplicación (ver figura 2) se encuentran cuatro elementos fundamentales: 1) *What?*, el cual se refiere al conjunto de datos que van a ser analizados; 2) *Who?*, referente a quién va a estar destinado el análisis; 3) *Why?*, relacionado al objetivo del análisis; 4) *How?,* concerniente al proceso que se aplica para lograrlo, y en este punto es donde se encuentra inmerso el proceso de aplicación previamente tratado. Así mismo, EDM establece: 1) ambiente educativo, enfocado al entorno de los datos; 2) datos en bruto, relacionado con los datos con los cuales se trabajará; 3) técnicas de minería de datos, respectivo a los procesos a aplicar; 4) interpretación de resultados, recomendaciones y divulgación, aplicado al cómo se utilizará el conocimiento descubierto (ver figura 3).

A partir de los elementos establecidos en EDM y LA, se definió el siguiente marco de trabajo, cada uno con las etapas y actividades necesarias para poder ejecutar los procesos de descubrimiento en el contexto educativo donde se ejecutó el proyecto de investigación (ver figura 4). Como se puede observar, se inició con el estudio del fenómeno de la deserción estudiantil partiendo de su revisión literaria en diferentes IES, para posteriormente centrar el análisis en la institución de educación superior sobre la cual se realizó el estudio.

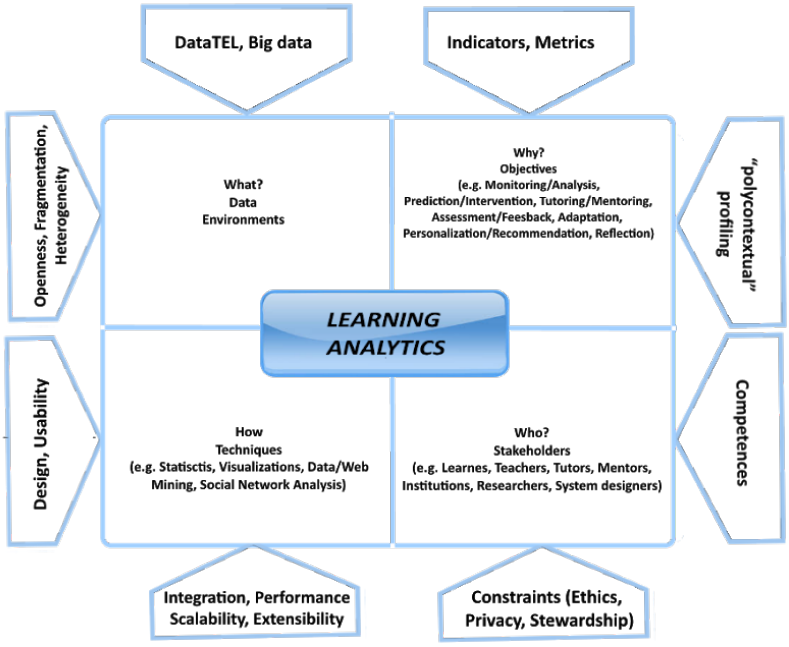

Figura 2. Modelo Learning Analytics
Fuente: A. Chatti, *et al*. [10].

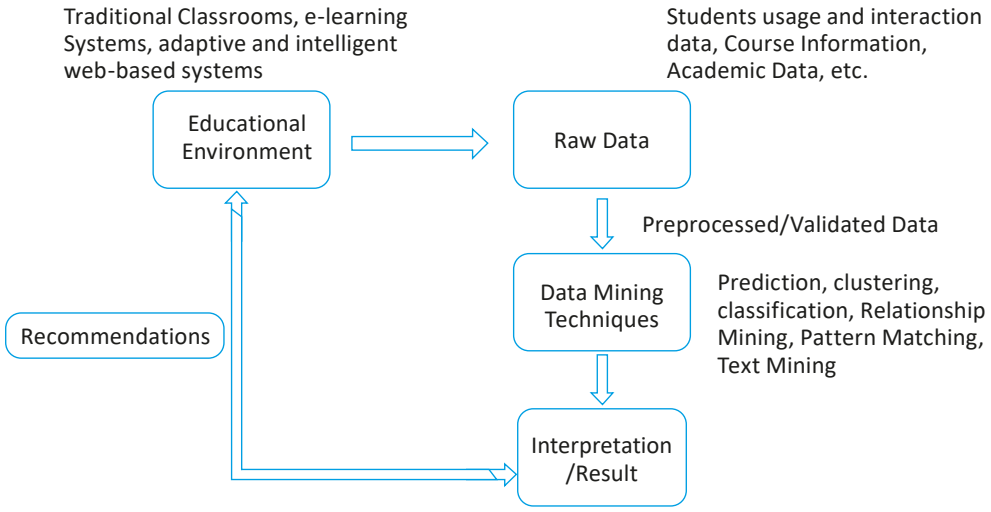

Figura 3. Proceso EDM (Educational Data Mining)
Fuente: Anoopkumar y Zubair [11].

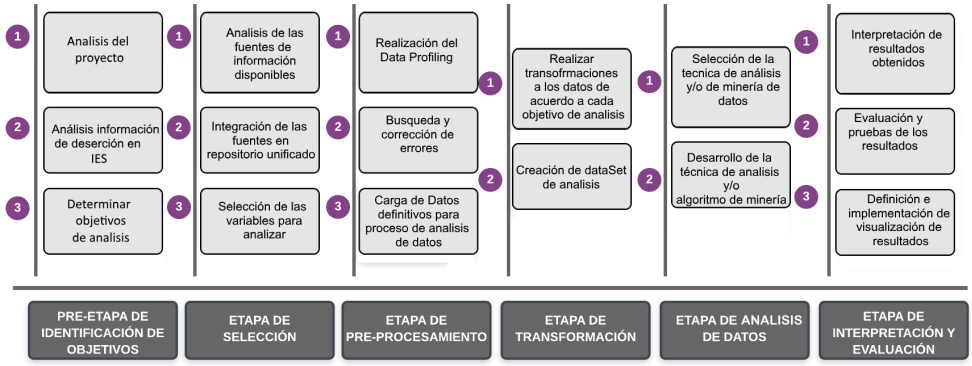

Figura 4. Metodología establecida a partir de LA – EDM
Fuente: elaboración propia.

Basados en este análisis, se procedió a identificar y seleccionar los datos y conceptos disponibles en las fuentes de la institución que relacionan características personales, socioeconómicas y académicas de los estudiantes, integrándolos en una bodega de información para su preprocesamiento y unificación para obtener un conjunto de datos depurados y listos para la aplicación de técnicas de minería de datos.

1.1 Pre-etapa de identificación de objetivos

Aquí se refiere el contexto de la problemática, una contextualización del fenómeno de deserción, revisión de literatura relacionada con la deserción estudiantil y la definición de criterios de deserción para este contexto en particular, determinando los objetivos del proceso de análisis de datos descritos a continuación:

- Identificar el perfil y características socioeconómicas de los estudiantes que se perfilan como desertores en el programa de Ingeniería de *Software*.
- Definir patrones socioeconómicos y académicos que permitan analizar el rendimiento académico de los estudiantes de Ingeniería de *Software* de la institución de educación superior analizada.

1.2 Etapa de selección

El objetivo de esta etapa es determinar las fuentes de datos y el tipo de información a utilizar, aquí la información relevante para el análisis es extraída de la fuente de datos seleccionando los referentes socioeconómicos de los estudiantes matriculados y sus respectivos registros de académicos, notas parciales de cada periodo académico obtenidas de la base de datos de la institución universitaria, recabando información desde el primer periodo académico del 2007 hasta el segundo semestre del 2017,

correspondiente a 634 registros de estudiantes con 51 atributos y el historial académico conformado por 13.282 registros, en los que se observan las tres notas parciales (primera y segunda con un peso del 30 % y la final 40 %) que por reglamento institucional se realizan en un semestre académico.

1.3 Etapa de pre-procesamiento

Los beneficios del conocimiento extraído dependen en gran medida de la calidad de datos tomados para el análisis, por ende, no solo es necesario haber realizado una buena recolección de fuentes y variables, sino que los datos deben estar completos e íntegros para obtener un análisis consistente con las necesidades. Es usual encontrar una variedad de errores en los datos almacenados y pueden darse por fallos al introducirlos, pérdida de información, diferencia de formatos, entre otros. Esta etapa tiene el objetivo de eliminar el mayor número de errores e inconsistencias dentro de las fuentes, de esta manera presentarlos de una forma apropiada para la etapa de análisis [12] .

Ahora, tras realizar un reconocimiento previo de cada una de las fuentes disponibles, se identificó una alta calidad de los datos, encontrando que solo un 3 % de los estudiantes presentaban errores de valores nulos en atributos sensibles como estado civil y socioeconómico, por lo que se corrigieron usando el valor modal o medio del campo, en la fuente de historial académico no se encontró inconsistencia en los atributos usados.

1.4 Etapa de transformación

A partir del pre-procesamiento realizado en la etapa anterior, el siguiente paso del proceso es la transformación de los datos. Esta engloba cualquier proceso que modifique o altere la forma de los datos, sin embargo y de acuerdo con la revisión de la literatura, existen procedimientos que trasforman un conjunto de atributos en otros, derivan en nuevos o bien cambian el tipo mediante numerización o discretización, o el rango mediante técnicas de escalado [12].

En este trabajo, al conjunto de datos usado se le realiza una discretización de atributos continuos con el fin de favorecer el análisis predictivo y descriptivo. En la tabla 1 se exponen ejemplos de la discretización realizada a campos de la fuente relacionada con los datos de los estudiantes.

1.5 Etapa de análisis de datos

Esta tiene como objetivo realizar un análisis intensivo para tratar de deducir o identificar patrones y tendencias existentes dentro de los datos, por lo general estos patrones o

tendencias no pueden ser deducidos por medio de consultas, esto se debe a la cantidad de registros y variables almacenadas.

Tabla 1. Ejemplo de discretizados en la fuente de estudiantes

Atributo	*Regla de discretización*	*Valor asignado*
Edad	Edad >=15 y edad <=20	1
	Edad >=21 y edad <=23	2
	Edad >=24 y edad <=28	3
	Edad >=29 y edad <=56	4
Régimen de Seguridad Social	Contributivo (de las EPS)	1
	Subsidiado (el del Sisbén)	2
Estado Civil	Casado(a)	1
	Madre soltera	2
	Separado	3
	Soltero (a)	4
	Unión libre	5
	Viudo(a)	6

Fuente: elaboración propia.

Para la realización de esta fase, se utiliza EDM en el abordaje del primer objetivo y LA como solución al segundo objetivo propuesto.

Para la realización de EDM, considerando que la recolección de información versaba sobre la predicción de un valor discreto (desertor: sí o no), se seleccionó la técnica *árbol de decisión* haciendo uso del algoritmo de clasificación C4.5, implementado en la herramienta *Orange* que se ha desarrollado en la Facultad de Informática de la Universidad de Ljubljana (Eslovenia). Este dispone de una programación visual para el análisis exploratorio de datos y visualización, contiene características potenciales de buena usabilidad, en la figura 5 se presentan los parámetros utilizados en la implementación del algoritmo.

Cabe resaltar que en las etapas iniciales de aplicación de LA en este proyecto, se realizó un proceso de revisión sistémica para determinar en una institución de educación superior colombiana, cuál sería el tipo de herramienta más adaptable y eficiente que podría aplicarse en un ámbito académico, dando como resultado la herramienta Power BI de Microsoft [13]. Posteriormente, se procedió a diseñar e implementar cada uno de los *Dashboards* que permitirían realizar análisis descriptivos sobre los datos [11].

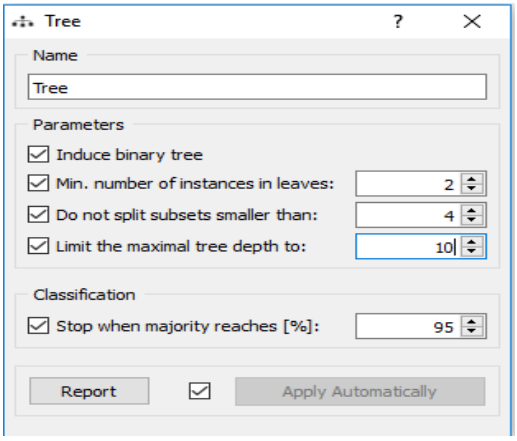

Figura 5. Parámetros usados para el árbol de decisión
fuente: elaboración propia.

2. RESULTADOS

2.1 Resultados *Educational Data Mining*

La fase de analítica de datos aplica técnicas predictivas utilizando el enfoque EDM y el algoritmo del *árbol de decisión*, dando como resultado lo expuesto en la figura 6, donde se representan los resultados en forma de árbol n-ario, pues de un nodo padre o raíz se desprenden *n* cantidad de hijos.

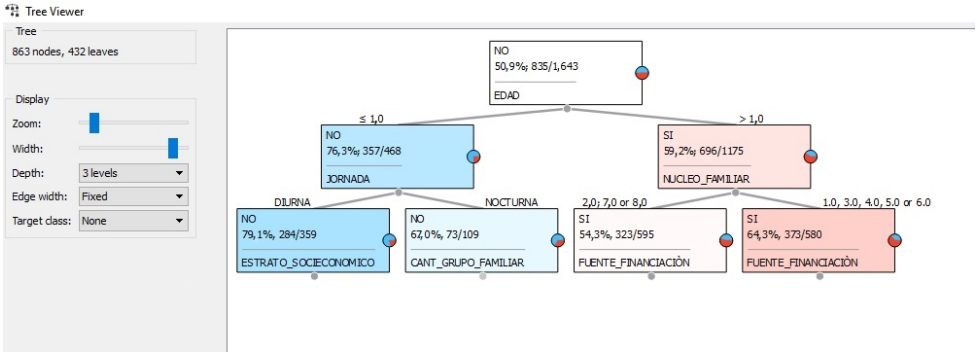

Figura 6. Visualizador del árbol de decisión generado por el algoritmo de minería
Fuente: Elaboración propia.

Aquí se maneja una escala e intensidad de color donde cada tono representa los posibles valores que puede tomar la variable, para este caso la variable *Desertor* en color rojo o *No desertor* en color azul, la intensidad del color depende de la concentración de la variable, cuando es más oscuro quiere decir que esta se presenta con mayor

frecuencia. Para la clase *No desertor* se obtuvieron como resultados más relevantes dos patrones con una probabilidad superior a 65 %, expuestos en las figuras 7 y 8, las cuales son expansiones del árbol general (ver figura 6) que presentan con mayor detalle el primer nodo allí expuesto.

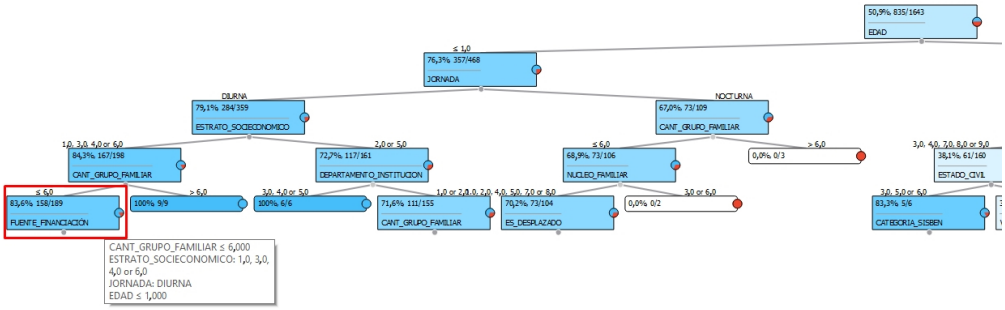

Figura 7. Patrón n.° 1 objetivo 1.1
Fuente: elaboración propia.

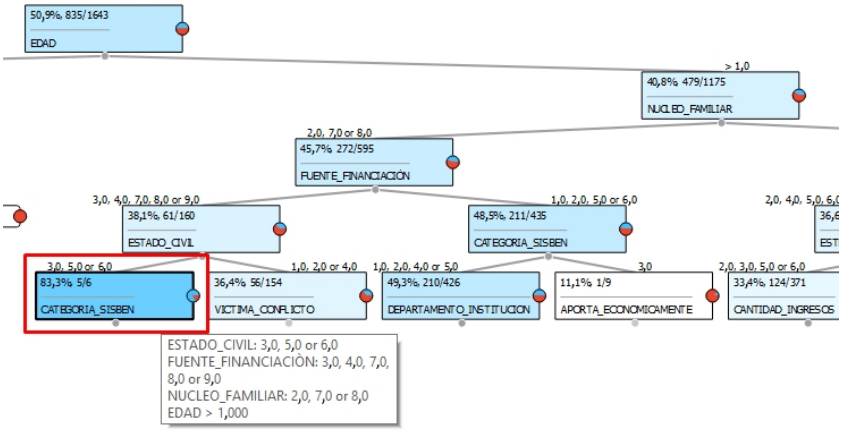

Figura 8. Patrón n.° 2 objetivo 1.1
Fuente: elaboración propia.

En el *árbol de decisión* correspondiente a la figura 7, resaltado en color rojo se puede observar el primer patrón encontrado en el cual los estudiantes cuya edad es menores o igual a 20 años (edad<=1), pertenecientes a la franja diurna con estrato socioeconómico diferente a 2 o 5 y cuyo grupo familiar este conformado por máximo seis integrantes tiene una probabilidad de 82,6 % de no desertar.

La figura 8 muestra resaltado en color rojo el nodo correspondiente que representa la segunda regla o patrón, indicando que el 83,3 % de los estudiantes no desertan del

programa académico si tienen una edad mayor a 20 años (edad>1), si viven con dos padres o tienen hijos, si la fuente de financiación no son recursos propios y su estado civil es separado, unión libre o viudo.

Por su parte, para los casos más relevantes en los que la clase es igual a SÍ (Sí desertar), se toma como referencia para el *árbol de decisión* el penúltimo nodo, generando a partir de él las imágenes expuestas en las figuras 9 y 10.

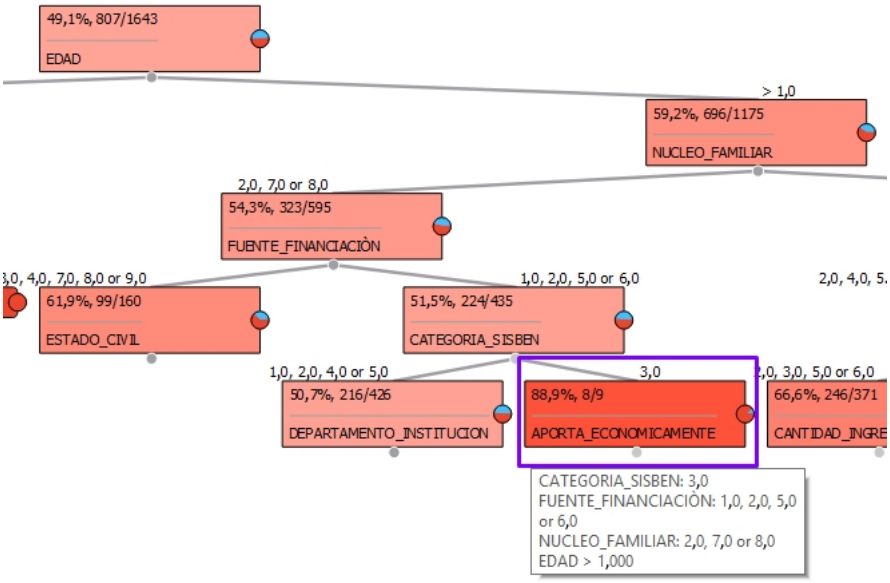

Figura 9. Patrón n.° 3 objetivo 1.1
Fuente: elaboración propia.

El *árbol de decisión* anterior (figura 9), corresponde al patrón número 3 relacionado con los estudiantes que desertan del programa académico, en este se observan los estudiantes con más de 20 años y cuyo núcleo familiar está compuesto por sus dos padres o sus hijos, y cuyos recursos para financiar los estudios no son propios, asimismo pertenecen a la categoría 3 del Sisbén con una probabilidad de desertar del 88,3 %.

En la figura 10 se resalta el nodo del *árbol de decisión* correspondiente al patrón número 4, este muestra que el 66,6 % de los estudiantes cuya edad es mayor a 20 años, que viven con alguno de sus padres o pareja y cuya fuente de financiación académica no proveen ellos o su pareja y pertenece a un estrato socioeconómico diferente de 1 y 4, desertan de sus estudios.

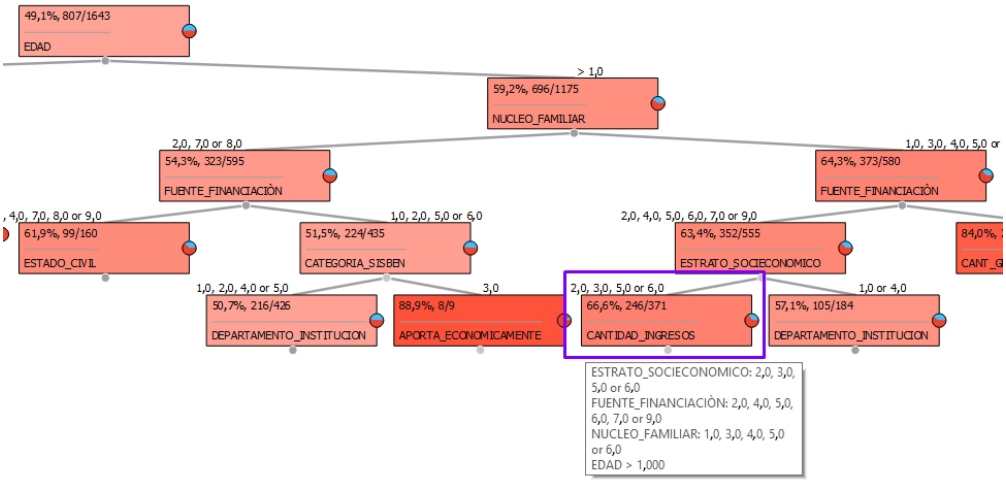

Figura 10. Patrón n.° 4 objetivo 1.1
Fuente: elaboración propia.

2.2 Resultados *learning analytics*

Con la herramienta seleccionada se diseñaron e implementaron diferentes *Dashboards* con el fin de realizar un análisis descriptivo de los datos, se realizó un proceso exploratorio con datos enfocados en el rendimiento académico de los estudiantes según: estrato social o discapacidad (haciendo uso del estrato socioeconómico, tipo de discapacidad, municipio, tipo de documento del estudiante, espacios académicos con su respectiva nota definitiva, promedio general de los estudiantes a través de los años, promedio general y georreferenciación según la dirección de residencia) (ver figura 11), situación laboral (con variables como dependencia económica, cargo, estado civil, género, estrato, notas de personas dependientes e independientes, notas por semestre y notas por espacio académico) (ver figura 12), jornada – nivel – semestre (jornada diurna–nocturna, nivel técnico, tecnológico, universitario, semestre académico, ciudad, programa, año, periodo, notas a través de los años, semestres y espacios académicos) (ver figura 13), y por núcleo familiar (número de hermanos, personas a cargo, vivienda propia, estado civil, tipo de documento, año cursado, estrato, georreferenciación por dirección y promedio de notas por espacio académico e historial a través de los años cursados) (ver figura 14), los cuales responden al segundo objetivo que plantea el proyecto.

Como se puede observar, es notoria la distribución socioeconómica de los estudiantes en el programa, siendo predominantes lo estratos bajos a medio–bajo (registrado a partir de la distribución del gráfico circular). También se identificó que los estudiantes con discapacidad física y motora bajo ciertas circunstancias, tienen un

mejor rendimiento académico que el promedio general (esto fue identificado a partir de la aplicación de filtros de búsqueda solo para personas con estas características, y se pudo observar un aumento notorio en las notas de toda la trayectoria académica frente al promedio general de los estudiantes).

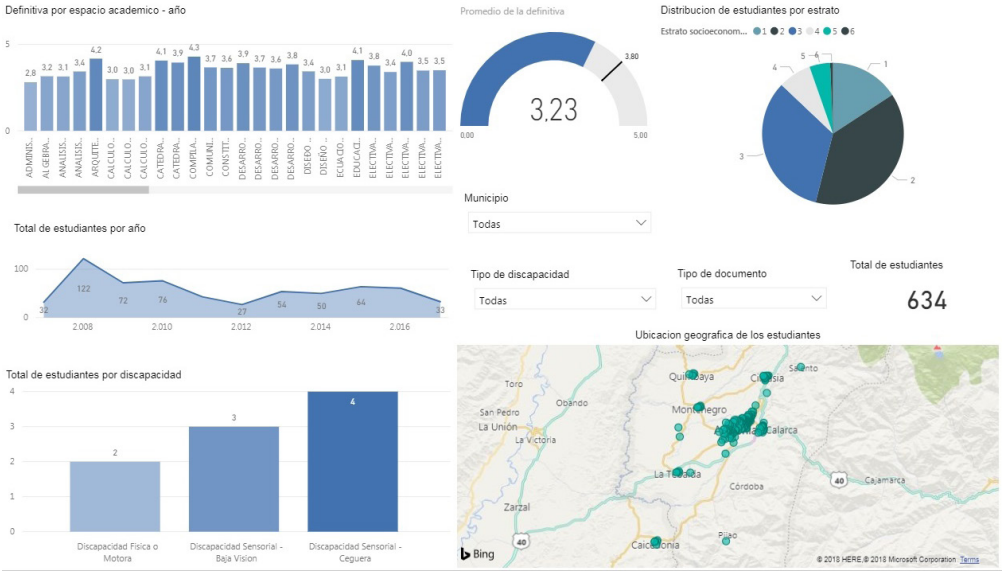

Figura 11. *Dashboard* estrato social o discapacidad
Fuente: elaboración propia.

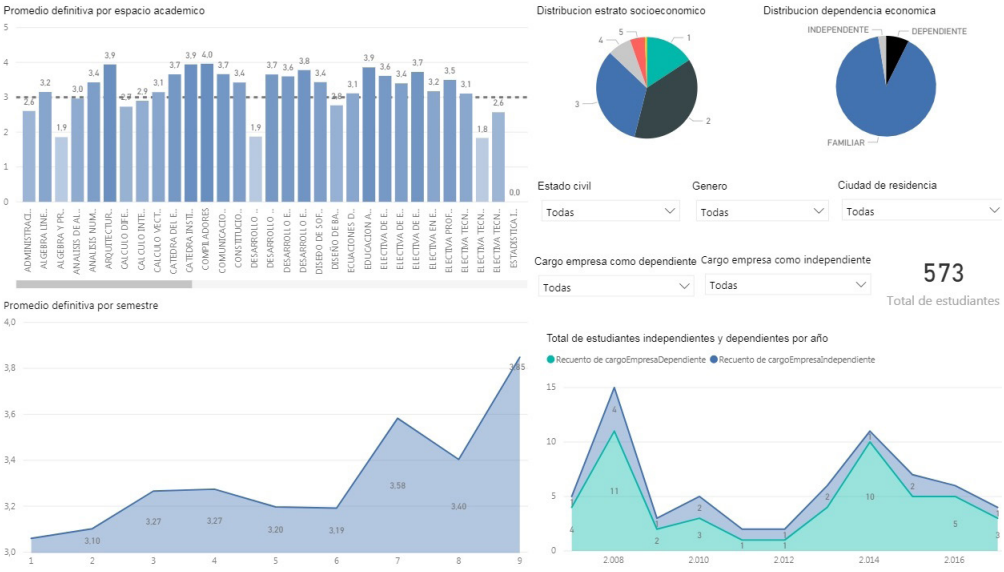

Figura 12. *Dashboard* situación laboral
Fuente: elaboración propia.

Cabe destacar que cada una de las gráficas utilizadas en los diferentes reportes y sus variables fue seleccionada a partir de las necesidades de los principales *stakeholders* en la institución universitaria donde se aplicó el proyecto. Con esto se identificó cada gráfica de barras, circular o torta, gráfico de líneas, mapas, filtros e indicadores claves de rendimiento, con el fin de enfocar los diferentes *Dashboards* de una manera amigable e intuitiva para que cualquier individuo de la comunidad académica pudiera hacer uso de este y entendiera fácilmente los resultados.

Con la aplicación de diferentes filtros sobre el *Dashboard*, se identificó que las mujeres tienen un mejor rendimiento académico, aunque es una carrera preferida por los hombres. También se identificó, con la aplicación de filtros, que solo una pequeña parte de la población estudiantil trabaja, por lo que tienen mejores habilidades comunicativas.

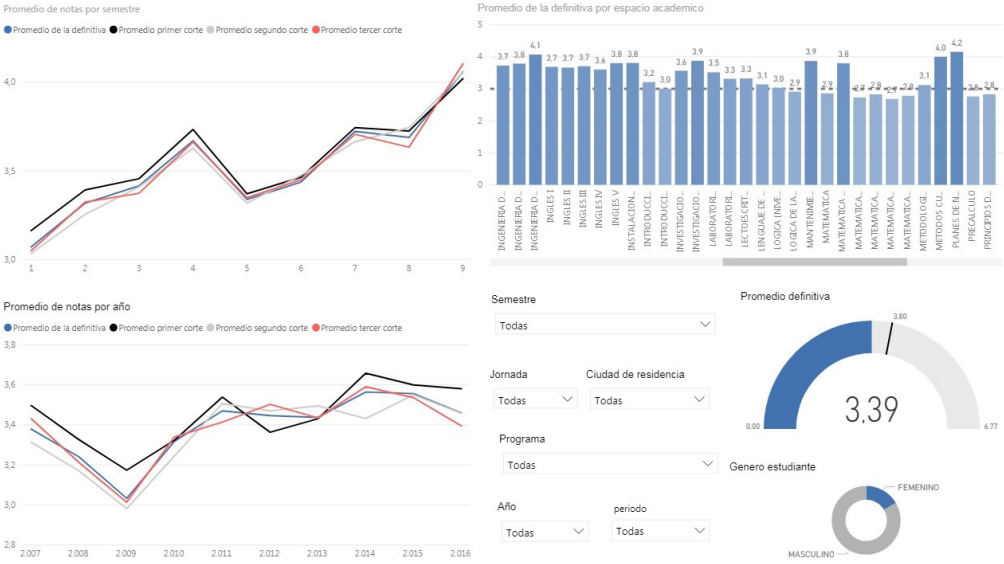

Figura 13. *Dashboard* jornada – nivel – semestre
Fuente: elaboración propia.

Como se puede observar en la imagen, es notoria la mejoría de los estudiantes a través de su crecimiento profesional, pero este cambia en el proceso transicional de técnico a tecnólogo (identificado a partir del cambio drástico del promedio general de los estudiantes en el gráfico de líneas de la parte superior derecha).

En este *Dashboard* fue posible identificar las materias donde los estudiantes tienen un mal rendimiento académico, las cuales demandan habilidades lógicas y de raciocinio. Además, se evidencia que la mitad de ellos son hijos únicos.

Figura 14. *Dashboard* núcleo familiar
Fuente: elaboración propia.

3. DISCUSIÓN

Respecto a la aplicación del proceso predictivo para responder las preguntas del enfoque EDM, se puede observar en la Figura 7 resaltado de color rojo el nodo cuya probabilidad de ocurrencia es de 82,6 %, además logran visualizarse las características necesarias (ramas o camino) que conllevan a este nodo, la ruta especificada es la siguiente; edad <=1 (estudiantes menores e iguales a 20 años) & Jornada= Diurna & (Estrato_Socioeconómico ≠ 2 o Estrato_Socioeconómico ≠ 5) & (Cant_Grup_Familiar <=6), es decir, aquellos estudiantes que cumplan con estas características tienen 82,6 % de probabilidades de no desertar, para este caso en concreto de 189 estudiantes que cumplen con esta regla 158 de ellos no desertaron. En la figura 8 se muestra resaltado en color rojo el nodo cuya probabilidad de ocurrencia es de 83,3 %, además se visualizan las características necesarias (ramas o camino) que conllevan a este nodo, la ruta especificada es la siguiente; edad >1 (estudiantes mayores de 20 años) & (Núcleo_Familiar = 2 (dos padres y hermanos) o Núcleo_Familiar = 7 (solo sus dos padres) o Núcleo_Familiar = 8 (sus hijos)) & (Fuente_Financiación ≠ 1 (becas de la Universidad) o Fuente_Financiación ≠ 2 (becas externas) o Fuente_Financiación ≠ 5 (Icetex) o Fuente_Financiación ≠ 6 (ingreso de padres o familia)) & (Estado_Civil =3 (Separado) o Estado_Civil =5 (Unión Libre) o Estado_Civil =6 (viudo(a)), es decir, aquellos estudiantes que cumplan con estas características tienen un 83,3 % de probabilidades de no desertar, para este caso en concreto de seis estudiantes que cumplen con esta regla cinco de ellos no desertaron.

En la figura 9 resaltado en color morado se observa el nodo cuya probabilidad de ocurrencia es de 88,3 %, además pueden visualizarse las características necesarias (ramas o camino) que conllevan a este nodo, la ruta especificada es la siguiente; edad >1 (estudiantes mayores de 20 años) & (Núcleo_Familiar = 2 (dos padres y hermanos) o Núcleo_Familiar = 7 (Solo sus dos padres) o Núcleo_Familiar = 8 (Sus hijos)) & (Fuente_Financiación = 1 (becas de la Universidad) o Fuente_Financiación = 2 (becas externas) o Fuente_Financiación = 5 (Icetex) o Fuente_Financiación = 6 (ingreso de padres y/o familia)) & Categoría_Sisbén = 3 (categoría 3), es decir, aquellos estudiantes que cumplan con estas características tienen un 88,3 % de probabilidades de desertar, para este caso en concreto de nueve estudiantes que cumplen esta regla 8 de ellos sí desertaron. En la figura 10 resaltado en color morado se puede evidenciar el nodo cuya probabilidad de ocurrencia es de 66,6 %, además pueden visualizarse las características necesarias (ramas o camino) que conllevan a este nodo, la ruta especificada es la siguiente; edad >1 (estudiantes mayores de 20 años) & (Núcleo_Familiar = 1 (alguno se sus padres) o Núcleo_Familiar = 3 (esposa (o compañeras y/o hijos) o Núcleo_Familiar = 5 (otros) o Núcleo_Familiar = 6 (solo)) & (Fuente_Financiación ≠ 1 (becas de la Universidad) o Fuente_Financiación ≠ 8 (usted y su esposa) & (Estrato_Socioeconómico ≠ 1 o Estrato_Socioeconómico ≠ 4), es decir, aquellos estudiantes que cumplan con estas características tienen 66,6 % de probabilidades de desertar, para este caso en concreto de 371 estudiantes que cumplen esta regla 246 de ellos sí desertaron.

Respecto al proceso exploratorio realizando en los *Dashboards* implementados para responder las preguntas del enfoque de *Learning Analytics,* se identificaron algunos comportamientos relevantes en los estudiantes del programa de Ingeniería de *Software*: 1) En general, las mujeres tienen un mejor rendimiento académico que los hombres en el transcurso de su formación como ingenieros de *software*, aunque es una carrera preferida por los hombres con un porcentaje de 83,44 % del total del alumnado; 2) se logra visualizar el mejoramiento constante de los estudiantes a través de su formación profesional, pero este comportamiento cambia en quinto semestre, esto puede deberse a que es el semestre donde se genera el cambio de formación de nivel técnico a tecnólogo y posteriormente a universitario (ver figura 13); 3) se logra evidenciar que los espacios académicos que tienen un promedio en la nota definitiva de los estudiantes por debajo de 3 son: Administración de bases de datos, Física I, Geometría, Lógica de programación, Matemáticas, Matemática discreta, Matemáticas aplicadas I, Matemáticas aplicadas II, Precálculo y principios de ingeniería de *software*, requiriendo estos espacios académicos alta demanda de habilidades lógicas; 4) se evidencia que alrededor del 72 % de los estudiantes son de estrato 2 y 3, además hay una clara relación entre el estrato socioeconómico y el rendimiento académico (ver figura 13), donde a mayor estrato socioeconómico más bajo es el rendimiento;

5) la mitad de los estudiantes del programa son hijos únicos; 6) solo el 10 % de los estudiantes tienen un trabajo de manera dependiente o independiente, el otro 90 % depende de su familia; 7) los estudiantes que trabajan como independientes, tienen mejores habilidades comunicativas y esto se ve reflejado claramente en las notas de los espacios académicos donde las habilidades de comunicación son necesarias, superando por mucho el promedio general; 8) los estudiantes con alguna discapacidad física o motora que se encuentran casados al momento de iniciar su proceso de formación profesional tienen un rendimiento académico superior al promedio académico general; 9) un factor representativo en la muestra analizada, es el impacto que tienen los estratos socioeconómicos en el rendimiento académico de los estudiantes y su posible deserción estudiantil. Por esta razón, se deben generar estrategias de concientización para fomentar y valorar el proceso de formación profesional en los estudiantes con el fin de nivelar estas diferencias entre estratos.

Un elemento que puede ser de interés para el aumento en el ingreso de los estudiantes al campo ingenieril del área de *software* es la acentuación de la difusión efectiva en los estratos más altos, debido a que estos no están teniendo una buena acogida, viéndose reflejado en la baja cantidad de estudiantes a mayor estrato socioeconómico.

La aplicación de EDM y LA en este proyecto fue enfocado a resultados diferentes: por un lado EDM fue utilizado en la identificación de elementos que permitieran observar patrones en la deserción estudiantil, mientras que LA, fue utilizado para realizar una descripción general sobre el rendimiento académico de los estudiantes según diferentes factores (familiar, laboral, socioeconómico, discapacidad y jornada-nivel-semestre), pero no se abordó la deserción, por lo tanto, no es posible contrastar los resultados obtenidos por ambos enfoques sobre un mismo conjunto de datos.

4. CONCLUSIONES

Este documento presenta los diferentes elementos que fueron tenidos en cuenta para la aplicación de LA y EDM en una institución de educación superior. Dichos análisis descriptivos se enfocaron principalmente en el rendimiento académico de los estudiantes y la deserción estudiantil a partir de datos estructurados, preprocesados, transformados, centralizados y analizados con las diferentes herramientas de minería de datos. Además, se presenta la posibilidad de integración de *Learning Analytics* con *Educational Data Mining* para su implementación en una institución de educación superior, los cuales son basados en KDD, esto permitiría encontrar conocimiento útil y novedoso que no es perceptible a simple vista, como patrones y tendencias en el rendimiento académico de los estudiantes y su deserción estudiantil. Adicionalmente, permite el manejo eficiente de la información y soporta la toma decisiones dentro de una

organización, brindando la posibilidad de mejorar aspectos o procesos de su entorno.

REFERENCIAS

[1] G. Siemens and R. Baker, "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration", *ACM. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia, Canada,* pp. 252-254, 2012.

[2] L. F. Cote Peña, "Hábeas Data en Colombia, un trasplante normativo para la protección de la dignidad y su correlación con la NTC/ISO/IEC 27001", Magíster en Derecho, Facultad de Derecho, División de Ciencias Jurídicas y Políticas, Universidad Santo Tomás Seccional Bucaramanga, Bucaramanga, Colombia, 2016.

[3] J. Vásquez Velásquez, E. Castaño Vélez, S. Gallón Gómez, and K. Gómez Portilla, "Determinantes de la desercion estudiantil en la Universidad de Antioquia", *Facultad de ciencias econimicas. Centro de investigaciones economicas, Universidad de Antioquia,* p. 43, 2003.

[4] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, *Advances in knowledge discovery and data mining.* Masachussetts: MIT Press, 1996.

[5] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases", *IA Magazine,* vol. 17, n.° 3, pp. 37-54, 1996.

[6] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, "The KDD process for extracting useful knoledge from volumes of data", *Communications of the ACM,* vol. 39, pp. 27-34, 1996.

[7] P. J. Goldstein and R. N. Katz, "Academic Analytics: The Uses of Management Information and Technology in Higher Education", *ECAR Research Study,* vol. 8, 2005.

[8] D. G. Oblinger and J. P. Campbell, "Academic Analytics", *EDUCAUSE White Paper,* 2007.

[9] D. Norris, L. Baer, J. Leonard, L. Pugliese, and P. Lefrere, "Action Analytics: Measuring and Improving Performance That Matters in Higher Education", *EDUCAUSE,* vol. 43, 2008.

[10] M. Amine Chatti, A. Lea Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics", *Int. J. Technology Enhanced Learning,* vol. 4, pp. 318-331, 2012.

[11] M. Anoopkumar and R. Zubair, "A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration", *IEEE,* pp. 1-12, 2016.

[12] J. H. Orallo, J. R. Quintana, and C. F. Ramirez, *Introducción a la mineria de datos.* Madrid, España: Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2005.

[13] J. A. Salazar Cardona and D. A. Angarita Garcia, "Evaluación y selección de herramientas de analítica visual para su implementación en una institución de educación superior", *IngEAM,* vol. 4, pp. 1-20, 2017.